EXHIBIT 10

12.3 A 1.2V 64Gb 341GB/s HBM2 Stacked DRAM with Spiral Point-to-Point TSV Structure and Improved Bank Group Data Control

Jin Hee Cho, Jihwan Kim, Woo Young Lee, Dong Uk Lee, Tae Kyun Kim, Heat Bit Park, Chunseok Jeong, Myeong-Jae Park, Seung Geun Baek, Seokwoo Choi, Byung Kuk Yoon, Young Jae Choi, Kyo Yun Lee, Daeyong Shim, Jonghoon Oh, Jinkook Kim, Seok-Hee Lee

SK hynix, Gyeonggi, Korea

With the recent increasing interest in big data and artificial intelligence, there is an emerging demand for high-performance memory system with large density and high data-bandwidth. However, conventional DIMM-type memory has difficulty achieving more than 50GB/s due to its limited pin count and signal integrity issues. High-bandwidth memory (HBM) DRAM, with TSV technology and wide IOs, is a prominent solution to this problem, but it still has many limitations: including power consumption and reliability. This paper presents a power-efficient structure of TSVs with reliability and a cost-effective HBM DRAM core architecture.

HBM's major obstacle for achieving high bandwidth and low power is the heavy capacitive load due to the thousands of TSVs in 8Hi stacks. The previous HBM (with multi-drop TSV) had limited capabilities for managing the 8Hi heavy TSV loading, which is one of the major obstacles for achieving both bandwidth and density. To reduce the heavy loading of a conventional multi-drop TSV structure, a spiral point-to-point (P2P) TSV structure is proposed. Figure 12.3.1 shows a conventional multi-drop TSV structure [1] and the proposed spiral P2P TSV structure. Compared to a multi-drop TSV structure, where each core die has its own TX, RX, and a 4-to-1 MUX, the proposed spiral P2P TSV structure has only three sets of TX and RX for an 8Hi stack. Furthermore, it eliminates the need for a channel selection MUX and its complicated routing. The current consumption to drive TSVs was reduced by 30%, due to the reduced capacitance, and its slew rate was increased from 3.4 to 4.9V/ns as shown in Fig. 12.3.1.

Another challenge for HBM development is achieving high yield for TSVs and micro-bumps (μ-bump): the total yield of a stacked chip is obtained by squaring the yield of one TSV by the number of TSVs. Furthermore, in case of an 8Hi-stack HBM, one failed TSV connection causes 9 dies to be discarded. Therefore, a TSV repair technique is essential to compensate for TSV yield. Unlike conventional TSV repair techniques [2] that need test equipment to test for the open/short-state of a TSV connection, the proposed automatic TSV self-repair technique, shown in Fig. 12.3.2, performs open/short tests during the boot-up sequence, without the need for test equipment or fuses. To detect a weak TSV connection, a core-side strong PMOS and a base-side NMOS, for which the leakage current is controlled by a bias voltage, are turned on at the same time. The quantized voltage level of the TSV is stored in a latch. A base-side strong PMOS and core-side NMOS are also turned on to confirm the TSV connectivity. By sequentially reading out the latched results, the locations of TSV failures can be determined. And the TSV self-repair operation can be performed by changing core die to find exact positions of failed TSVs. The slice ID signal sent from the base die to the core die is changed using the test mode to make each core die behave like the top slice. The proposed architecture also supports a conventional current scan, without additional circuitry, by jointly using the PMOS' as a current source and the DFF as a switch-enable signal shifter. The TSVs used to control the repair operations cannot make use of self-repair, but instead they exist in pairs to ensure robust operation. The repair procedure is performed during the boot-up sequence, so that users do not need to execute post-package repair.

HBM2 has a pseudo-channel function [1], which decreases the page size in half while doubling the effective number of banks to improve DRAM core timing such as t_{FAW} (four-active-window) and t_{CCDL} (column-to-column access timing). Prior work [1] used four channels in a slice, whereas the proposed architecture includes only two channels per core die via a spiral TSV structure. Each channel is divided into two pseudo channels: a pseudo channel has 16 independent banks and 64 IOs. Because of the large number of IO lines, the area overhead of HBM is significant, thus leveraging increased pre-fetch, which is a common method in conventional DRAM, is restricted. The improved bank group control, shown in Fig. 12.3.3, is proposed to mitigate speed and area penalties. For a 4b pre-fetch operation, each IO has 4 internal data lines, and there should be 64×4b×4 bank group IOs (BG_IO) and 256 global IOs (GIO) per each pseudo channel. However, the proposed architecture has BG_IOs with a 2b pre-fetch (feasible due to a relaxed t_{CCDL}) and GIO has a 4b pre-fetch, that is divided into left and right, to keep the effective line similar to the 128 GIOs. Figure 12.3.3 shows the timing diagram of the proposed architecture where column commands and Y-pulses are divided into even and odd groups by the order of the command input (BL0, 1 and BL2, 3). Therefore, the core die effectively has twice the number of data lines and column addresses, and it has twice the core timing margin without significant area penalty.

The power distribution in the HBM core die is also a challenge because significant power is consumed in a small area. The IR drop in a 3D stack structure causes t_{CDDL} degradation in the DRAM core operation. To mitigate this issue, additional bank-power TSVs are placed in-between row decoders: these directly supply power from base die to power-hungry core areas. Another merit of bank power TSVs is the ability to share the power distribution network among each core die, which greatly reduces the peak IR drop for an 8HI-stack as shown in Fig. 12.3.4. Based on PDN simulation results, more than 50% of IR drops are diminished in IDD4W (gapless write; worst pattern for core IR drop) compared to a previous version without bank power TSVs.

The base die and core dies in HBM2 all have temperature sensors. Memory controllers can read out the 8b of highest temperature code among the core dies, 8b for the base die temperature code, and 1b of catastrophic trip threshold (CATTRIP). Therefore, 4Hi and 8Hi stacks require 36 and 72 TSVs. The proposed serial temperature read-out scheme uses only 2 TSVs, one for the temperature code and the other for the CATTRIP. Figure 12.3.5 shows the core die temperature read-out scheme. The core die temperature codes are shifted, in descending order of core dies, by the shift clock that is generated in base die. The base die stores 8b of code in the CTEMP register, and compares it with the code stored in the CTEMP_MAX register, which stores the maximum temperature code observed. The CATTRIP scheme is depicted in the right side of Fig. 12.3.5. Since CATTRIP must reflect all information from every core die and base die, all dies share one TSV using wired-OR logic. The base die always turns on a pull-down transistor to drive the CATTRIP μ-bump LOW. When any die reaches the limit temperature (e.g. 125°C), it generates a CATTRIP flag to make the TSV HIGH, which is driven onto the CATTRIP μ-bump.

Since the PHY μ-bump cannot be probed, all tests of the HBM were performed using a direct access ball (DA). However, because of the operational characteristics of the PHY IO and the necessity for system implementation verification, an active interposer package (AIP), depicted in Fig. 12.3.6, is proposed. HBM DA and PHY operation can be verified, and 2-channel interleaving technology can be applied to determine the influence of the independent operations between channels. The AC characteristics of the HBM PHY can be measured: such as the input setup/hold, the 1-pin input setup/hold, and t_{DV} (data valid window). Since the signal integrity between the controller and the HBM in the SiP is reflected on the AIP, the LFSR/MISR can be operated between the active interposer and the HBM under similar conditions. Test flexibility is improved by applying a serial-test-mode input technique similar to IEEE1500 [3].

Several key technologies have been introduced to address impediments to increasing bandwidth for HBM memories. The spiral-P2P scheme and the TSV self-repair scheme both provide good solutions for managing the heavy 8Hi TSV loading. The improved bank-group data control optimizes area overhead and DRAM core speed. Additional bank power TSVs reduce the IR drop at the bank side by 50%. The HBM shmoo results, shown in Fig. 12.3.6, shows a 341GB/s 8Hi known-good-stack dies (KGSD) gapless-read operation at 1.2V and 105°C, and 320GB/s at 1.15V and 105°C. The chip micrographs for the 8Hi-stacked 8Gb DRAM dies and the base die are shown in Fig. 12.3.7.

References:

[1] J. C. Lee, et al., "A 1.2V 64Gb 8-channel 256GB/s HBM DRAM with peripheral-base-die architecture and small-swing technique on heavy load interface," *ISSCC*, pp. 318-319, 2016.
[2] D. U. Lee, et al., "An Exact Measurement and Repair Circuit of TSV Connections for 128GB/s High-Bandwidth Memory(HBM) Stacked DRAM," *IEEE Symp. VLSI Circuits*, 2014.
[3] JEDEC Standard High Bandwidth Memory (HBM) DRAM Specification, 2015.

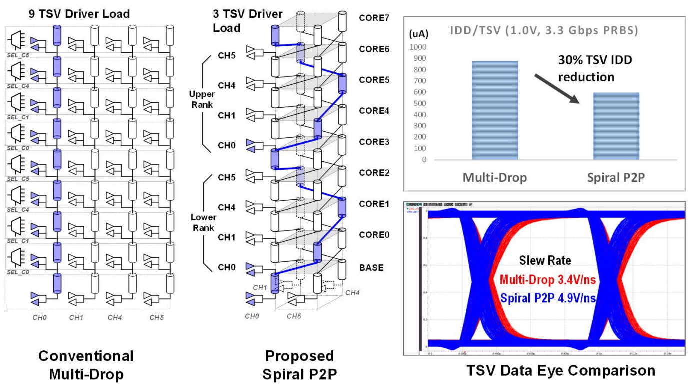


Figure 12.3.1: Structure and performance comparison between multi-drop and the spiral P2P TSV structure.

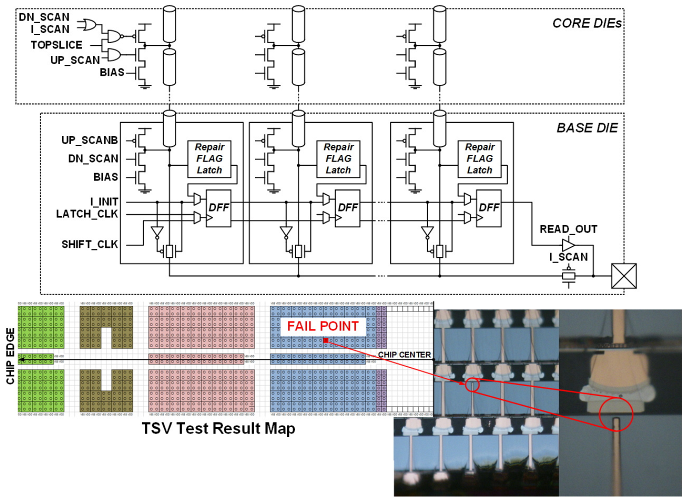


Figure 12.3.2: TSV self-repair scheme and test results.

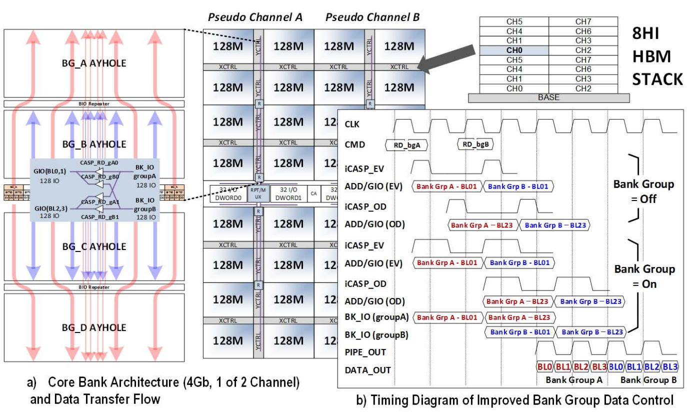


Figure 12.3.3: Core architecture and improved bank-group data control.

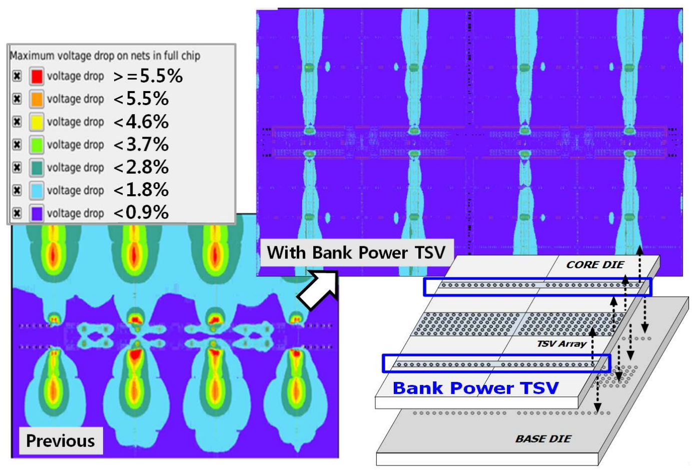


Figure 12.3.4: Power TSVs in the middle of banks and power distribution network simulation results.

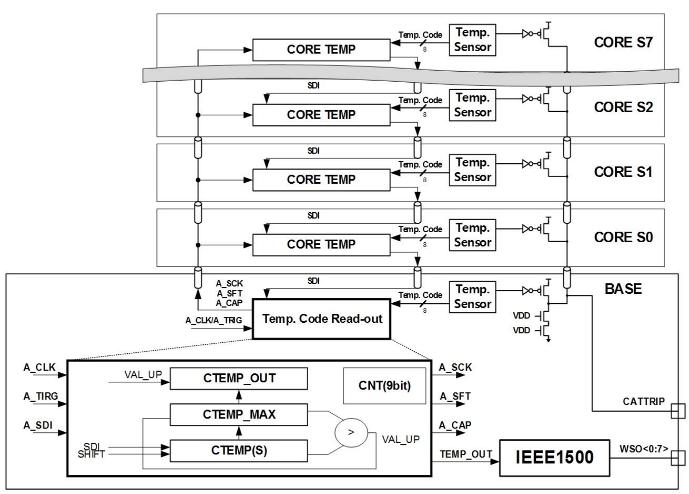


Figure 12.3.5: Serial temperature read-out and the catastrophic trip threshold scheme.

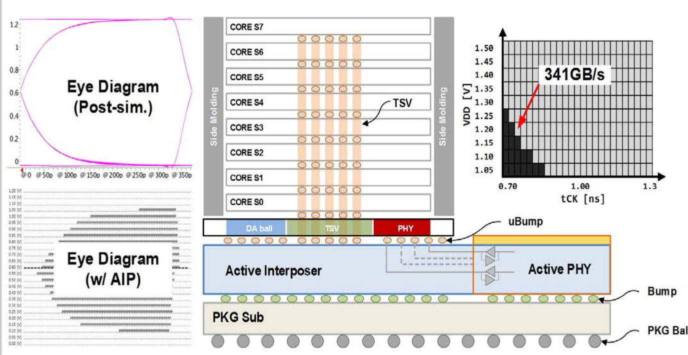
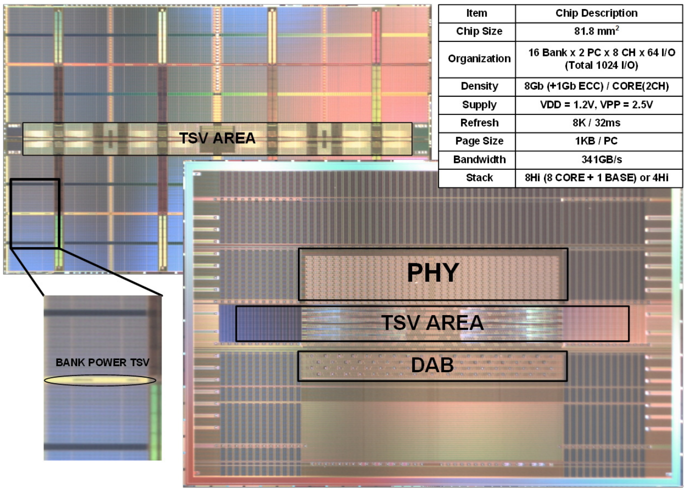


Figure 12.3.6: Active interposer package and test shmoo result.

12



Item	Chip Description
Chip Size	81.8 mm^2
Organization	16 Bank x 2 PC x 8 CH x 64 I/O (Total 1024 I/O)
Density	8Gb (+1Gb ECC) / CORE(2CH)
Supply	VDD = 1.2V, VPP = 2.5V
Refresh	8K / 32ms
Page Size	1KB / PC
Bandwidth	341GB/s
Stack	8Hi (8 CORE + 1 BASE) or 4Hi

Figure 12.3.7: Chip micrograph and summary table.